ProxiViz: an Interactive Visualization Technique to Overcome Multidimensional Scaling Artifacts

Nicolas Heulot∗ CEA, LIST, 91191 Gif-sur-Yvette CEDEX, France
Michaël Aupetit† CEA, LIST, 91191 Gif-sur-Yvette CEDEX, France
Jean-Daniel Fekete‡ INRIA Saclay

ABSTRACT
Projection algorithms such as multidimensional scaling are often used to visualize high-dimensional data. However, when attempting to interpret the visualization of the resulting 2D projection, users are faced with artifacts. This poster introduces ProxiViz: an interactive technique to provide better insights about the original data-space. Primary results of a controlled experiment show that ProxiViz is significantly more effective than the baseline projection techniques for a visual clustering task.

Index Terms: H.5.2 [Information Interfaces and Presentation]: User Interfaces

1 INTRODUCTION
A large number of visualization techniques have been designed to reveal the information hidden in multidimensional data. A popular method in the context of exploratory analysis consists in projecting data in a two- or three-dimensional space and visualize it as a scatterplot. However the resulting display is often biased by artifacts of dimensionality reduction. These artifacts imply challenges of interpretation and trust [4] for the analysts who want to make inferences about data, such as outliers detection or visual clustering.

There are also several possible mechanisms to convey more information than the bare projection, such as feedback on the quality of the projection. This poster introduces the visualization technique ProxiViz which uses a color encoding of the original distance information coupled with interactions to help analysts exploring high-dimensional data using projections. We also discuss the technique considering the different kinds of artifacts.

Figure 1: The 3D sphere on the center is projected favoring two different artifacts: left exhibits a sphere projected using PCA with false neighborhoods (items far away in data-space are close-by in low-dimensional space) and right exhibits an unfolded sphere with tears (items close-by in data-space are far away in low-dimensional space). An example of ProxiViz is presented below each projection.

Figure 2: Proximity-based visualization with a color encoding of the Voronoi cells (left) and interpolated color encoding (right), using the proposed color scale (the white dot indicates the reference point).

2 RELATED WORK
In multidimensional projection, data instances \( p_i \in P \) are characterized by a set of attributes that defines their multiple dimensions. We call this space the data space. Many projection algorithms exist to map the data-space to a low-dimensional space [3]. This transformation, called dimensionality reduction, often introduces artifacts, making the resulting projection noisy and difficult to interpret.

There are many different visualization techniques and metrics to evaluate the quality of a mapping [5]. Quality measures or stress measures are mostly based on the definition of a criteria which compare the distances \( d_{i,j}^* \in D^* \) in the data-space with the distances \( d_{i,j} \in D \) in the low-dimensional space. A common technique used with linear mappings displays the stress measure, optimized by the projection algorithm, as a color on each point using a color scale from yellow to red, as suggested by Bentley and Ward [2].

Aupetit [1] propose to visualize the distances \( d_{i,j}^* \) in the data-space relative to a reference point \( s \), using a color encoding of the Voronoi cell of each point. In this proximity-based visualization, the reference is interactively selected by the user on the projection visualization. Artifacts are indirectly visualized by contrast between positions in \( D \) and colors representing distances in \( D^* \).

Because the stress measures are not rich enough to ensure visual inference of the clustering, contrary to the proximity-based visualization, we decided to focus on this last interactive technique and improve it.

3 DESIGN
The original proximity-based visualization uses a grey color scale and a coloring of the Voronoi cells. The Voronoi cell around a point \( x \) being the set of all pixels whose distance to \( x \) is not greater than their distance to the other points. We introduce here a new color encoding and interactions to better deal with the multidimensional scaling artifacts.

3.1 Color encoding
We propose to use the color scale adapted for visual comparison tasks [7]. This color scale allows to better differentiate the borders between clusters, as it helps to segment the distances between intra-cluster and inter-cluster distances. It starts with a white color indicating the null distance \( d_{\text{max}}^* \) in the data-space, then continues with purple and green nuances, and ends with black (Fig. 2), indicating the maximum distance \( d_{\text{max}}^* \) to the reference in the data-space.

Figure 2: Proximity-based visualization with a color encoding of the Voronoi cells (left) and interpolated color encoding (right), using the proposed color scale (the white dot indicates the reference point).
Instead of coloring the Voronoi cells, we propose to use an interpolated coloring. Indeed the size of the Voronoi cells does not represent any information and then create a bias that surprises users. We propose to use the Shepard interpolation [6] between the points. This interpolation computes the color \( u(x) \) of a pixel \( x \) using the inverse distance weighting of the color \( w_i \) of each point \( i \) on the projection visualization: 
\[
u(x) = \sum_{i=0}^{N} w_i \frac{|x - x_i|}{\sum_{i=0}^{N} |x - x_i|} \]

We choose a neighborhood factor \( k = 2 \) to better preserve the local information. This interpolation displays the global distance trends in each area of the projection visualization while preserving the local distance information with a shaded circle around each point.

### 3.2 Interactions

When the mouse moves over the visualization area, the nearest point is interactively selected as the reference data instance and the color encoding of the visualization is immediately updated to display the distances from the data-space.

However, this implementation requires some care: while moving the mouse to explore the visualization, the color map changes rapidly when hovering over close-by points that are actually far away in the data-space (false neighborhoods). This flicker, which is a witness of the artifact, is also disturbing to users while exploring the visualization, the color map changes rapidly when hovering over close-by points that are actually far away in the data-space (false neighborhoods). This flicker, which is a witness of the artifact, is also disturbing to users while exploring the cluster (see Fig. 3).

![Figure 3: Example of flickering problem due to false neighborhoods: three clusters were projected closer to each other in low dimensional space than they are actually in the data-space. Therefore, while switching between reference points close from each other on the projection, the color mappings are completely different (the largest dot indicates the reference point).](image)

To address this issue, we implemented an interaction technique that overcomes the brutal switch. Using the user feedbacks from a preliminary evaluation of the technique, we identified two main user exploration strategies:

- **Cluster exploration**: users want to explore a specific cluster, avoiding false neighbors flickers but speeding-up over non connected parts of the cluster on the projection (tears).

- **Free exploration**: users want to explore the whole projection, not a specific cluster.

To take these two strategies into account, we propose the following interaction technique: when the mouse pointer moves away from the reference point \( x \) used to display the distances, closer to point \( j \), we do not select this new point as the reference immediately. We set a timer instead with the timeout condition as
\[
T = T_0 \times d_{i,j}^3 \quad \text{with} \quad T_0 \text{ a time constant (e.g. 600 ms per high-dimensional distance unit).}
\]

If the mouse pointer comes closer to another new point before the timeout, then the timer is canceled and a new one is set. Otherwise, the new point used to set up the timeout is selected as current reference point and the distances are displayed.

This technique allows to explore a cluster without interferences caused by false neighborhoods. It also allows to explore a cluster splitted by tears without selecting a reference point from another cluster between the two or more disjoint components of the splitted cluster.

We call ProxiViz this new design of the proximity-based visualization technique, using color interpolation and augmented with the interaction technique described above which allows a fluid exploration of the projection.

### 4 Discussion

ProxiViz with its stabilized interaction helps to smoothly explore a projection. However the type of projection algorithm used to map the data-space to the low-dimensional space is also important. A projection algorithm, which tend to preserve locally in low-dimensional spaces the neighborhood that exist in high-dimensional space will avoid false neighborhoods but will be prone to tears artifacts (see Fig. 1). Conversely, several popular projections such as PCA will produce false neighborhoods. ProxiViz allows to deal with both of these artifacts, but we observed that it was easier to make sense of projections mainly composed of tears than of false neighborhoods, even with ProxiViz.

We performed a controlled experiment of ProxiViz compared to baseline projection techniques, considering a task of visual clustering. The participants were asked to count clusters on different projections of several datasets. The statistical results showed in particular that the users were significantly more accurate using ProxiViz than the bare projection.

### 5 Conclusion

We recommend the use of the interactive ProxiViz technique, a proximity-based visualization technique designed to easily explore a projected multidimensional data-space using an interpolated color encoding of the data-space distances. We believe ProxiViz can greatly help users make sense of multidimensional projections, which are theoretically effective but practically not widely used due to their high level of abstraction from data to display. ProxiViz uses colors to convey distances but we intend to explore other methods, such as the transient displacement of point positions during interactive exploration of the projection visualization. We hope more research will help for the adoption of multidimensional projections techniques to facilitate the exploration of high-dimensional spaces.

### Acknowledgements

This research has been funded by CEA and DGA.

### References


